Various Sampling Techniques

- **Random sampling**
  - Every unit of the population has the same probability of being included in the sample.
  - A chance mechanism is used in the selection process.
  - Eliminates bias in the selection process
  - Also known as probability sampling
- **Nonrandom Sampling**
  - Every unit of the population does not have the same probability of being included in the sample.
  - Open the selection bias
  - Not appropriate data collection methods for most statistical methods
  - Also known as non-probability sampling

"Data Science & AI" by Siva Rama Krishna

Random Sampling Techniques

- Simple Random Sample
- Stratified Random Sample
  - Proportionate
  - Disproportionate
- Systematic Random Sample
- Cluster (or Area) Sampling

Simple Random Samples

- Every object in the population has an equal chance of being selected
- Objects are selected independently
- Samples can be obtained from a table of random numbers or computer random number generators
- A simple random sample is the ideal against which other sample methods are compared

"Data Science & AI" by Siva Rama Krishna

Stratified Random Sample

- Population is divided into non-overlapping subpopulations called strata
- A random sample is selected from each stratum
- Potential for reducing sampling error
- Proportionate -- the percentage of these sample taken from each stratum is proportionate to the percentage that each stratum is within the population
- Disproportionate -- proportions of the strata within the sample are different than the proportions of the strata within the population

Systematic Sampling

- Convenient and relatively easy to administer
- Population elements are an ordered sequence (at least, conceptually).
- The first sample element is selected randomly from the first *k* population elements.
- Thereafter, sample elements are selected at a constant interval, *k*, from the ordered sequence frame.

$$k = \frac{N}{n},$$

where:

n = sample size

N = population size

k = size of selection interval

"Data Science & AI" by Siva Rama Krishna

Cluster Sampling

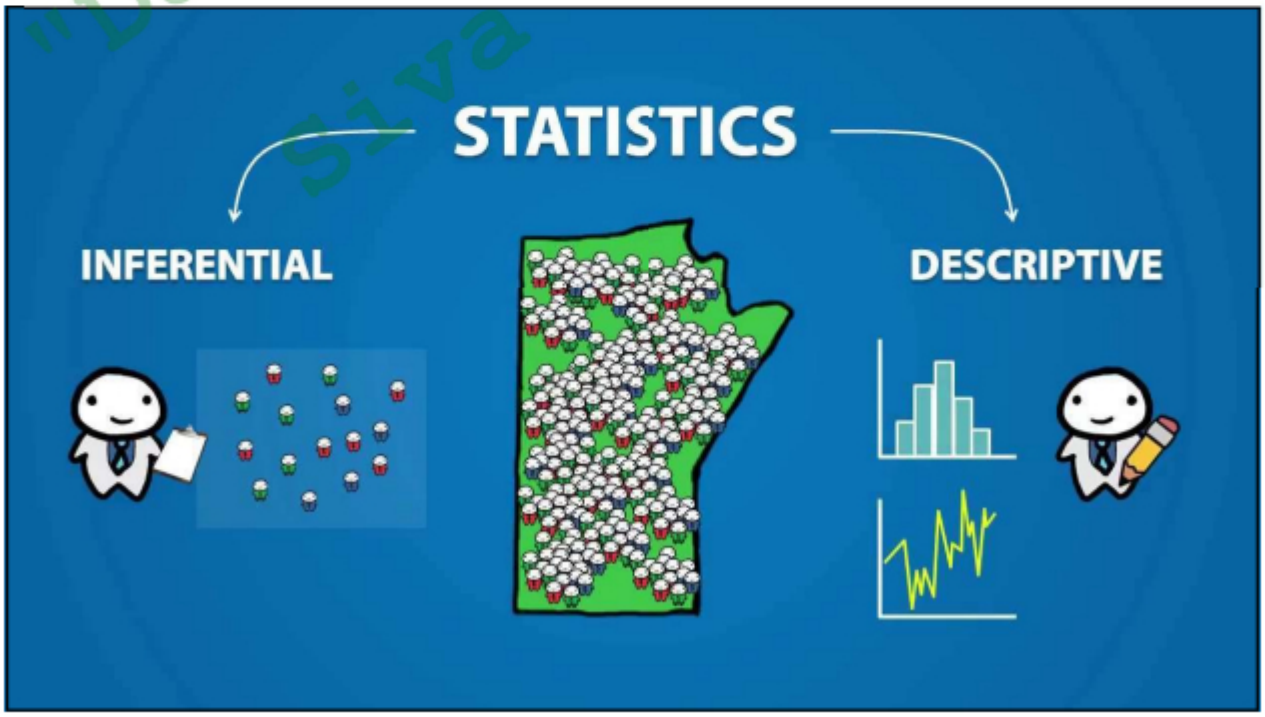
- Population is divided into non-overlapping clusters or areas
- Each cluster is a miniature of the population.
- A subset of the clusters is selected randomly for the sample.
- If the number of elements in the subset of clusters is larger than the desired value of *n*, these clusters may be subdivided to form a new set of clusters and subjected to a random selection process.

Nonrandom Sampling

- **Convenience Sampling:** Sample elements are selected for the convenience of the researcher
- **Judgment Sampling:** Sample elements are selected by the judgment of the researcher
- **Quota Sampling:** Sample elements are selected until the quota controls are satisfied
- **Snowball Sampling:** Survey subjects are selected based on referral from other survey respondents

"Data Science & AI" by Siva Rama Krishna

"Data Science & AI" by Siva Rama Krishna



Descriptive vs Inferential Statistics

- **Descriptive statistics**
  - Collecting, presenting, and describing data
- **Inferential statistics**
  - Drawing conclusions and/or making decisions concerning a population based only on sample data

"Data Science & AI" by Siva Rama Krishna

Populations and Samples

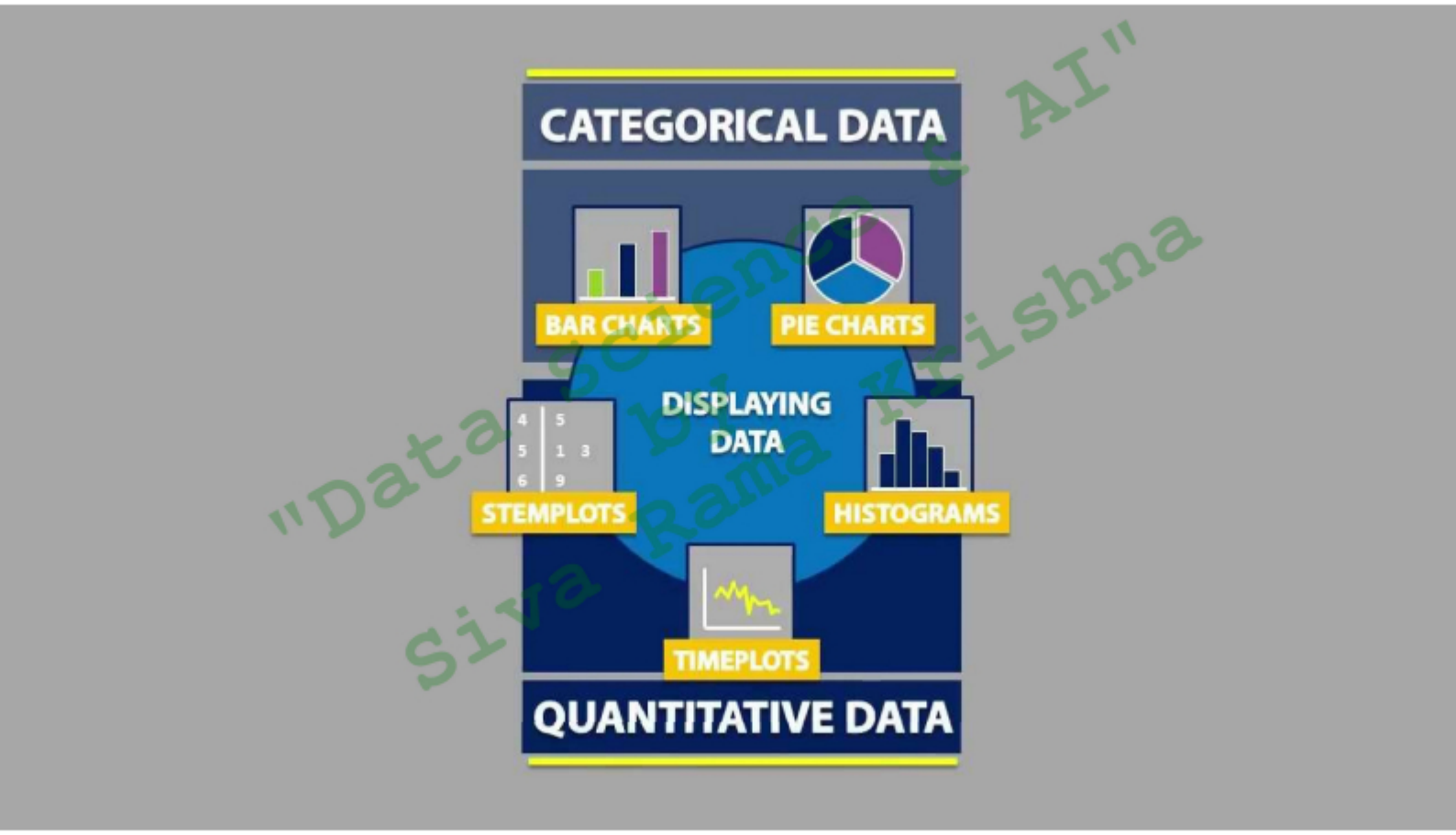
- A **Population** is the set of all items or individuals of interest
  - Examples: All likely voters in the next election
    All parts produced today
    All sales receipts for November
- A **Sample** is a subset of the population
  - Examples: 1000 voters selected at random for interview
    A few parts selected for destructive testing
    Random receipts selected for audit

"Data Science & AI" by Siva Rama Krishna

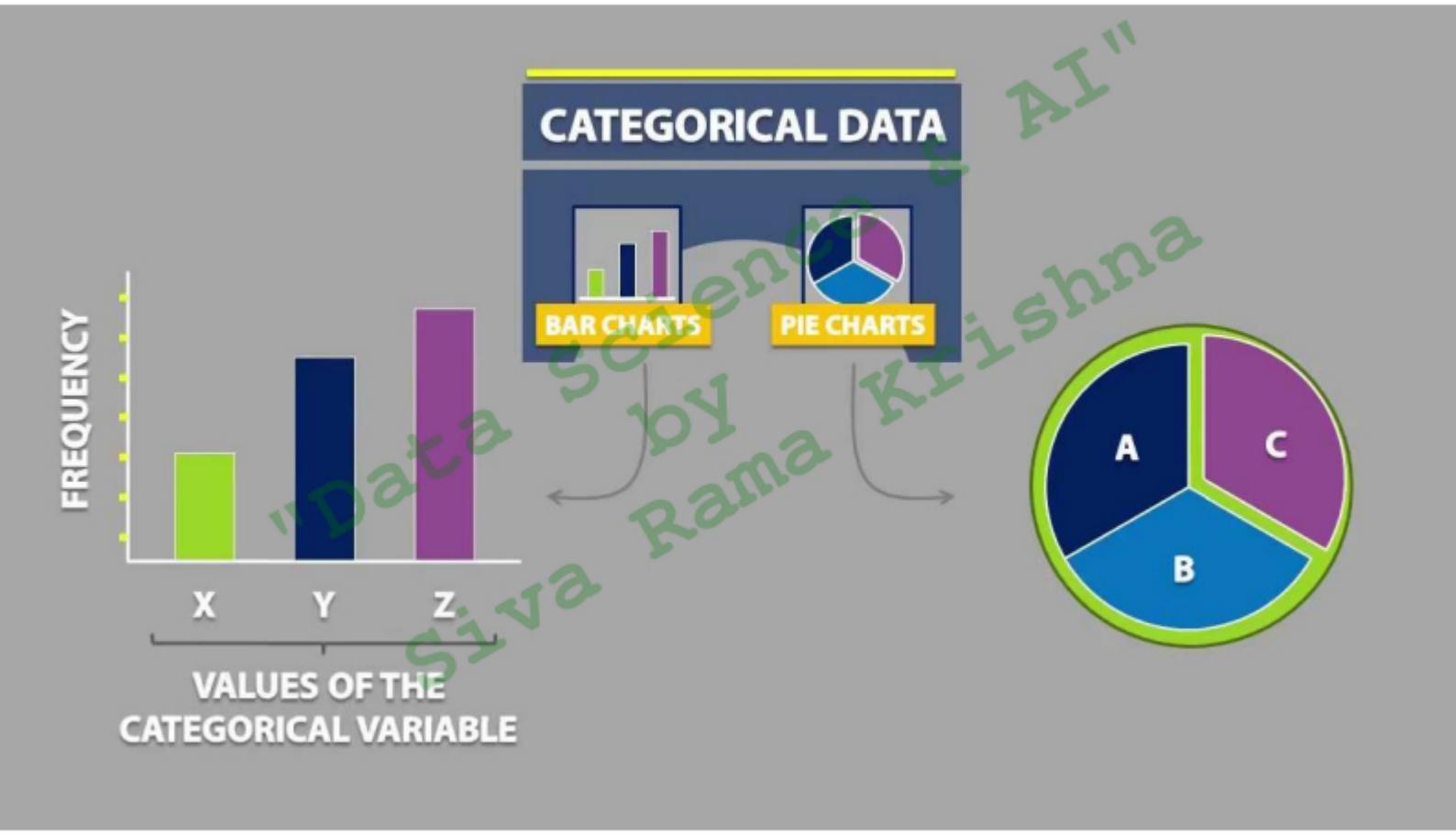
Why Sample?

- Less time consuming than a census
- Less costly to administer than a census
- It is possible to obtain statistical results of a sufficiently high precision based on samples.
- Because the research process is sometimes destructive, the sample can save product
- If accessing the population is impossible; sampling is the only option

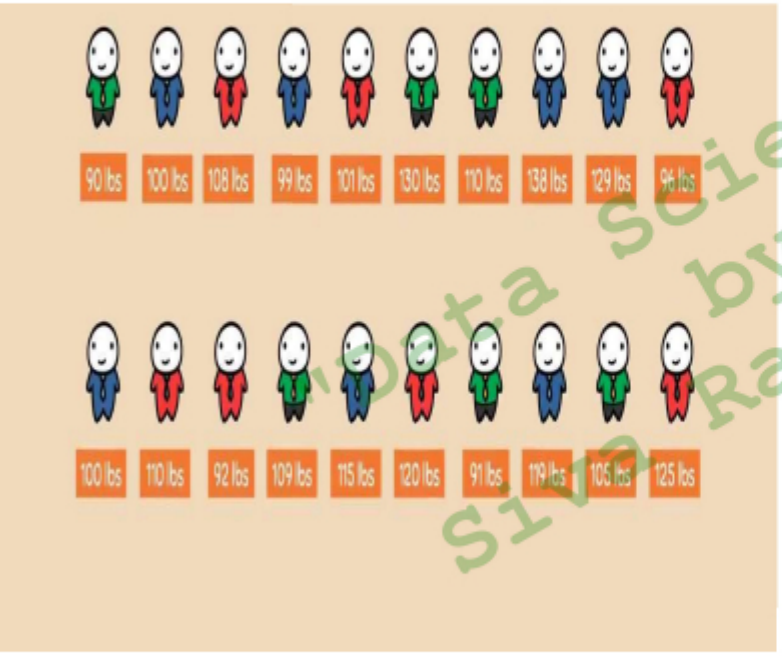
"Data Science & AI" by Siva Rama Krishna

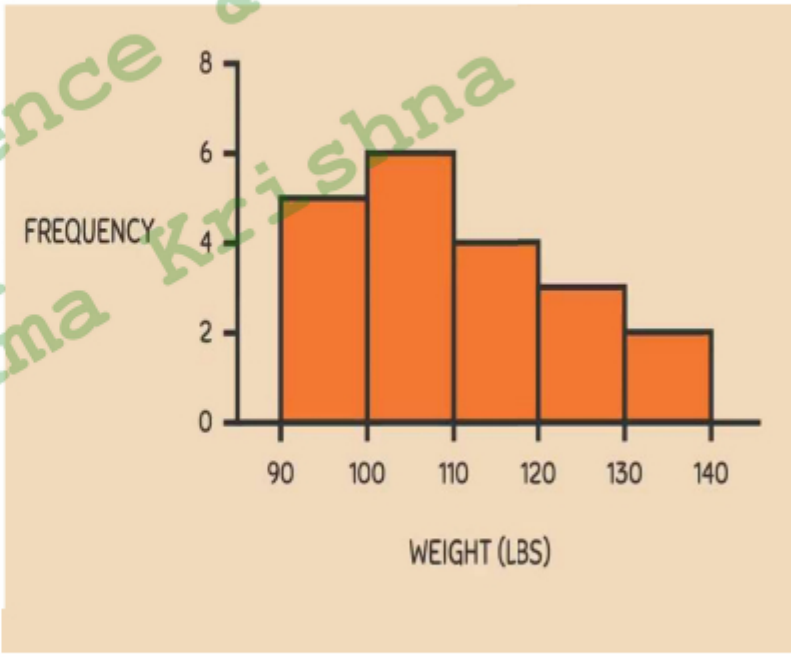


"Data science & AI"
Siva Rama Krishna



"Data Science & AI" by Siva Rama Krishna





"Data Science & AI" by Siva Rama Krishna

FREQUENCY DISTRIBUTION

WEIGHT	
100 – 110	
110 – 120	120
120 – 130	130
130 – 140	
140 – 150	

BY CONVENTION, WE SAY THAT EACH INTERVAL DOES NOT INCLUDE THE RIGHT END POINT

Dept. Science & IT" Siva Rama Krishna

TIMEPLOT

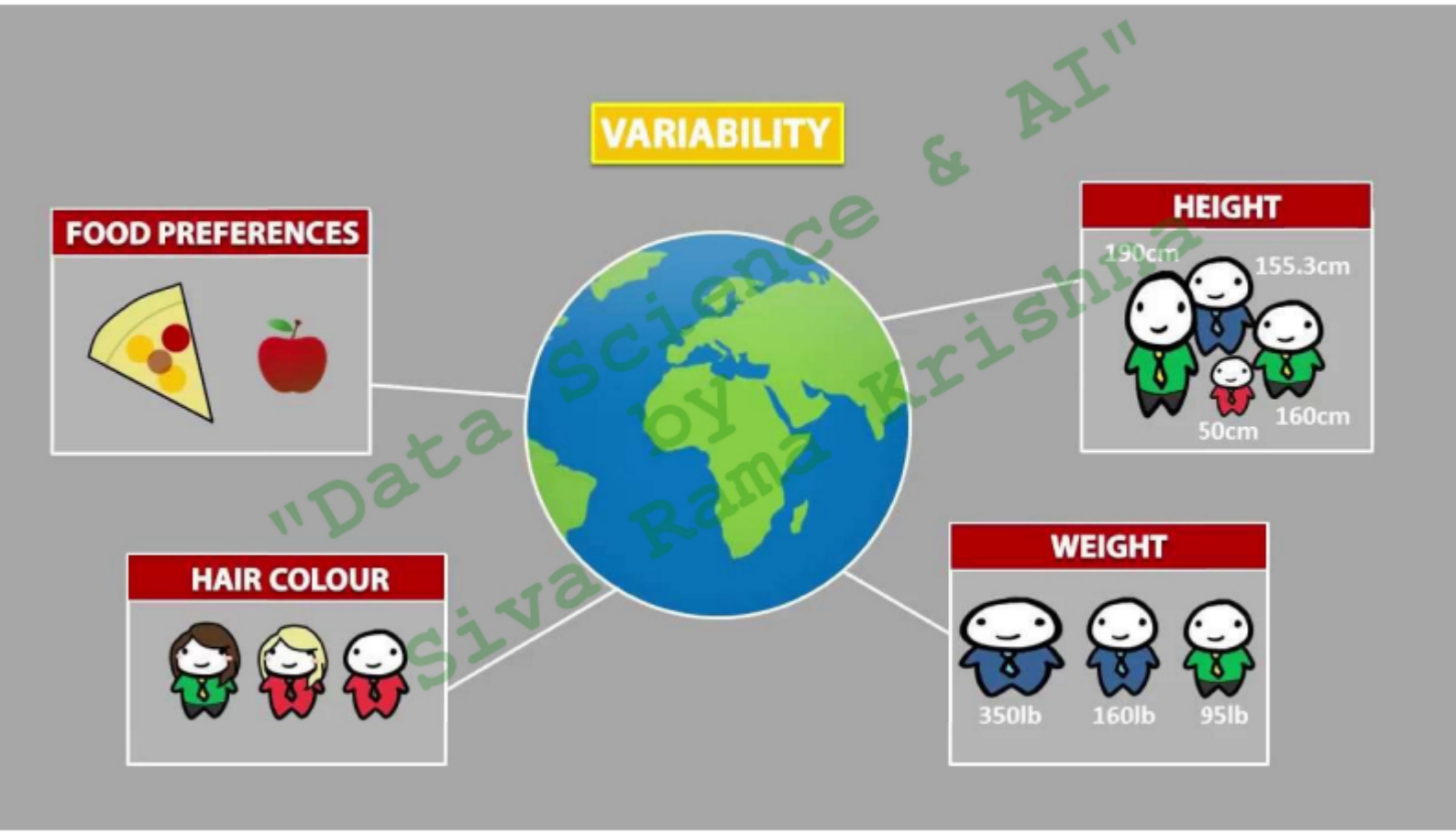
SHOW HOW A VARIABLE CHANGES OVER TIME

"Data Science & AI"
Siva Rama Krishna

TIMEPLOT

SHOW HOW A VARIABLE CHANGES OVER TIME

"Data Science & AI" by Siva Rama Krishna



"Data Science & AI" by Siva Rama Krishna

MEASURE

VARIABLE

QUANTITATIVE DATA

DATA THAT IS MEASURED IN NUMBERS. IT DEALS WITH NUMBERS THAT MAKE SENSE TO PERFORM ARITHMETIC CALCULATIONS WITH

QUANTITATIVE VARIABLES

HEIGHT

WEIGHT

MIDTERM SCORE

CATEGORICAL DATA

REFERS TO THE VALUES THAT PLACE "THINGS" INTO DIFFERENT GROUPS OR CATEGORIES

CATEGORICAL VARIABLES

HAIR COLOUR

TYPE OF CAT

LETTER GRADE

Data Science & AI" by Siva Rama Krishna

CATEGORICAL VARIABLE

CATEGORICAL AND ORDINAL

LOGICAL ORDERING TO THE VALUES OF A CATEGORICAL VARIABLE

EX: LETTER GRADE

F C C+ B B+ A A+

CATEGORICAL AND NOMINAL

NO LOGICAL ORDERING TO THE VALUES OF A CATEGORICAL VARIABLE

EX: HAIR COLOUR

RED BLONDE BROWN BLUE

"Data Science & AI" by Siva Rama Krishna

QUANTITATIVE VARIABLE

DISCRETE

REFER TO VARIABLES THAT CAN ONLY BE MEASURED IN CERTAIN NUMBERS

EX: NUMBER OF PETS YOU OWN

0 1 2 30 2.7

CONTINUOUS

REFER TO VARIABLES THAT CAN TAKE ON ANY NUMERICAL VALUE

EX: WEIGHT

105 185 170.683

"Data Science & AI" by Siva Rama Krishna